International Journal of Information Retrieval Research
Volume 11 • Issue 1 • January-March 2021

An Approach to Extract New Keywords From Radical Groups in Social Networks

Geetika Sarna, Netaji Subhas University of Technology, India
M. P. S. Bhatia, Netaji Subhas University of Technology, India

ABSTRACT

In recent times, numerous users as well as communities on social networks post messages in multimedia formats. The significant part of the message is the keyword that would help in recognizing the theme of information. Hence, this research aims to determine the new keywords occur in the messages posted on social network which would also be beneficial in identifying the category of user, various communities, and hidden patterns exist in the social network. In this paper, probabilistic approach is applied to identify the new keywords from the radical groups. Radical groups are those whose demeanor is totally opposite to the acceptance of community, for instance, terrorist groups. Hence, the dataset of terrorist community extracted from Twitter is used to find the new keywords that occur for a short span of time. State-of-the-art studies carried out the identification of terrorist communities based on keywords already present in lexicon, but the proposed approach makes the decision on the basis of both old as well as new keywords.

KEYWORDS

Bayes Rule, Overlap Community, Probability, Radical Groups, Social Network, Terrorist Community

1. INTRODUCTION

Today enormous amount of communities are available on the social network that may include sports, education, music, terrorist etc. These communities have been generated based upon the common interest. Out of these existing communities, some are beneficial for mankind and some believe in the destruction of mankind. The users of the particular community share common interest and communicate with each other on the common topic. For instance: sports community discusses the events related to the sports using the keywords related to the sports theme. Hence, these keywords would become an essential parameter in identifying the domain or the community to which the user belongs. With the advancement in the technology, it would become easy for the people to be a part of social network through any digital media such as smart phone, laptop, tablets etc. Consequently, large number of users have been joining the social network that may result in the production of large number of new keywords on the social network. Since new keywords are appearing frequently, so it is essential to find the way to keep track on new words to keep the dictionary up-to date in order to get more accurate results.

DOI: 10.4018/IJIRR.2021010103

In this paper, we are focusing on the radical groups as they are very harmful for the humanity of the whole world. Radical group means the group who wants to recognize enormous transformation to society according to their logics because they contemplate that the things are not favoring them. Being radical is not a bad thing if it is positive change, like the women's liberation movement of the 1960's and 1970's but it sometimes be the incipient of behaviour that inclines towards extremism and terrorism. Today terrorism has reached to its ultimate degree of terror. It becomes threat for the mankind. The formation of the threat is not possible for the single man but is created through a community or group. They are taking the help of online communication media to spread the terror or to accomplish their transgressed ideas against the whole world. They form communities on social media and exchange detrimental information which needs to be stopped. The main noteworthy point is that they mostly communicate using some keywords which are new to the dictionary. Those keywords may either be a code word or normal word. The time span of these new keywords is very small as they are only interested in sending and receiving the important messages for small duration in order to hide their identity. Here, our main focus is the identification of those new words which occur for that short span and are detriment to the society. Also, the linkage of the new keywords with other community is also detected in order to find its impact.

Existing work only considered the existing lexicon to identify the category of message but there is no research that could focus on the classification of the messages with the aid of new keywords also. This paper will focus on the extraction of newly appeared keywords in messages and use those keywords for the classification of messages more accurately. Also, the communities for new keywords will be identified and overlap between those communities will also be analyzed. The concept of overlapped communities will give the idea regarding the type or existence of connections or linkage of the user with other communities.

The remaining paper is structured as under. Related work is given in Section 2 followed by the proposed method to solve the problem in section 3. Results are given in Section 4. In Section 5, discussion about the overlap community is given and finally, in section 6, conclusion and future course of work is given.

2. RELATED WORK

2.1 Related Work on Keyword Extraction

Turney, P(1999) proposed Genetic algorithm named GenEx that automatically extract phrases from text data. Frank, E. et al. (1999) used two features: TFIDF(Term frequency Inverse document frequency) and distance of phrase from the beginning of the document in order to automatically extract key phrases from the text. Naïve Bayes classifier has been used as both the features are independent.Euler,T. (2002) proposed a method to extract sentences related to a certain topic on the basis of existing list of words. Yih, W et.al (2006) used features such as term frequency of each potential keyword, inverse document frequency, and frequency of the term occurs in search query logs to extract keywords from the advertisements posted on web pages..

Li, Z et. Al (2010) took the help of features such as linguistic feature, information retrieval feature, length of social snippet, document frequency and capitalization to extract keywords from the messages posted on Facebook. Bhowmik, R.(2010) proposed a method for the extraction of keywords from abstracts and titles of academic research articles using the Perceptron Training Rule . Wang, J et al.(2010), proposed a method for the extraction of Chinese keywords on the basis of association between words. Lee, K et al.(2013) proposed the extraction framework for the campaign that includes promotional and advertising campaigns, spam messages and political astroturfing. This framework employed three modules: (i) isolation of coherent campaigns from large graph of messages using graph mining techniques (ii) analysis of the comparison between text-based message correlation and user levels (iii) analysis of temporal behaviours of various campaign types. Nugumanova, A

Fuzzy Opinion: Detection of Opinion Based on SentiWordNet Dictionary by Using Fuzzy Logic